Supplementary Note 1.

Stage 1: Finding human α-helical transmembrane domains

The consensus human membrane proteome was previously predicted by using different transmembrane α-helix prediction methods[3-8], including TMHMM 2.0[1] (http://proteinatlas.org)[2]. However, the assembly of the human genome (ENSEMBL 36.52) has been updated since these studies were performed. Because our goal is to provide the basis for a dynamical up-to-date assessment, reflecting new membrane protein structures as they are determined, we used TMHMM 2.0 to identify transmembrane spanning proteins in an updated RefSeq version (BUILD 37.2) of the human genome[9]. For robustness of membrane protein identification, alignment, and clustering, only membrane proteins that are predicted to have at least two transmembrane α-helices were included in the analysis. For each full-length sequence, only the transmembrane domain from the first to the last predicted transmembrane α-helix residue was used for all subsequent analyses. To remove highly similar sequences from our analysis, the resulting domain set was clustered at 98% sequence identity (USEARCH[10]) followed by retaining only one representative for each cluster.

Stage 2: Assessing current modeling coverage

All transmembrane domain sequences were attempted to be modeled using ModPipe[11-13], which relies on PSI-BLAST[14] and MODELLER[15] for its functionality. Sequence-structure matches were established using fold assignment methods, including sequence-sequence[16], profile-sequences[14,17], and profile-profile[17,18] alignments, using a PDB[19] template database from 9/30/2011. The probability of finding a template structure was increased by using alignments with the E-value threshold of 1.0. By default, ten models were calculated for each alignment. A representative model for each alignment was then chosen by ranking based on the atomic distance-dependent statistical potential DOPE[20].

Nature Structural & Molecular Biology: doi:10.1038/nsmb.2508

Finally, the fold of each comparative model was evaluated using several quality scores, using an approach originally developed for globular protein structures[12]. A comparative protein structure model was considered 'reliable' if it scored better than certain threshold for at least one of the quality criteria (zDOPE[20] < 0, MPQS[12] > 1.0, GA341[21] > 0.7, TSVMod[22] native overlap > 0.4, target-template sequence identity > 30%) or was generated from a significant sequence-structure alignment (an alignment is considered significant, if the corresponding E-value is lower than 0.0001)[12]; reliable models are generally predicted to have the correct fold. The reliable models were deposited in the ModBase database of modeled structures (salilab.org/modbase/search?dataset=tmh_sequences). They provide a useful resource for investigating a membrane protein whose structure has not yet been determined by experiment.

The 'modeling coverage' was calculated for each modeled domain sequence, taking into account all reliable models produced by ModPipe, as follows. Each residue in the sequence was first annotated with the sequence identity between the modeled sequence and its closest template. The modeling coverage was then computed as the percentage of residues above a given sequence identity threshold. We also classified the modeling coverage of a sequence as 'high' (> 90% of domain residues are modeled), 'medium' (60-90%), and low (<60%).

Stage 3: Clustering of domain sequences

For assistance in target selection, transmembrane domain sequences were clustered into the smallest possible number of clusters, such that the structure of any member of a cluster allows for comparative modeling of the remaining cluster members at specified thresholds on target-template sequence identity and fraction of aligned residues (coverage)[23]. To obtain a pragmatic solution, we clustered the human domain sequences using BLASTCLUST[24] at several sequence identity (20%, 25%, 30%) and coverage (50%, 70%, 90%) thresholds. The BLASTCLUST algorithm includes in a cluster all domains that match at least

Nature Structural & Molecular Biology: doi:10.1038/nsmb.2508

one cluster domain at the specified sequence identity and coverage thresholds. A domain can only be a member of one cluster. As an aside, we also tested another algorithm, USEARCH[10], but the corresponding clusters were judged to be less suitable for the purposes of structural genomics (data not shown).

Stage 4: Assessing target selection strategies

We tested two target selection strategies, 'guided' and 'random.' These strategies were assessed with respect to how many target structures need to be determined by experiment to achieve varying degrees of structural characterization of the human membrane proteome by comparative modeling. Guided target selection prioritizes experimental structure determination of domains by the size of the clusters that they make accessible to modeling. In contrast, random target selection picks targets randomly from a list of domain sequences without known structures. In addition, the utility of the existing target lists of the nine PSI membrane protein centers (TargetDB[25]; October 2011) for modeling the human membrane proteome was computed; these targets were matched to the human α-helical transmembrane domains by pairwise sequence comparison using BLASTP[14].

Stage 5: Expanding the target set by adding homologous sequences

The cluster analysis was focused on the human sequences only, in the expectation that each human cluster will have a large number of non-human sequences related to at least one human member at more than 30% sequence identity. To get these non-human homologs, a multiple sequence profile of each human domain sequence was prepared by scanning the sequence against all 18.5 million sequences in the UniProt database (release-2011_08) using the BUILD_PROFILE module of MODELLER[18], as implemented in ModPipe (http://salilab.org/modpipe), first using the default settings (E-value: 0.1, five iterations) against a non-redundant (at the 90% sequence identity level) version of UniProt. The profiles were then expanded by one additional iteration using the

Nature Structural & Molecular Biology: doi:10.1038/nsmb.2508

full UniProt database. Once the profile was calculated, only homologs with the sequence identity of at least 30% were retained. The set of such homologs for all human domains in the cluster corresponds to potential structural genomics targets that allow modeling of any cluster member.

Nature Structural & Molecular Biology: doi:10.1038/nsmb.2508

Supplementary Note 2.

Stage 1: Finding human α-helical transmembrane domains

Of the 29,375 unique human protein sequences from the RefSeq-37 database of the human genome[9], 7,299 were predicted by the TMHMM 2.0[1] program to contain at least one transmembrane α-helix; 3,838 were predicted to contain two or more such helices. For each full-length sequence, only the transmembrane domain from the first to the last predicted transmembrane α-helix residue was used for all subsequent analyses. Only a single representative of sequences with more than 98% sequence identity to each other was retained, using program USEARCH[10], yielding the final non-redundant dataset of 2,925 domains (Suppl. Fig. 1a).

Stage 2: Assessing current modeling coverage

To provide context for assessing various target selection lists and strategies, we assessed the current ability to model each of the human α-helical transmembrane domain sequences. Automated comparative modeling using ModPipe[12] resulted in “reliable” models (Supp. Note 1) for 10-100% of residues in 2,683 of the 2,925 non-redundant human α-helical transmembrane domains. The modeling coverage of a domain sequence was calculated as the percentage of its residues that were modeled based on a template structure with the sequence identity to the modeled sequence above a given sequence identity threshold (Suppl. Fig. 1b). The modeling coverage was described as high, medium, and low when >90%, 60-90%, and < 60% of the domain residues, respectively, were modeled.

Nature Structural & Molecular Biology: doi:10.1038/nsmb.2508

Stage 3: Clustering of domain sequences

A cluster suitable for structural genomics target selection needs to satisfy two requirements to ensure that a structure determined for one of its members allows the modeling of most if not all other cluster members: (i) member sequences must be sufficiently similar to each other and (ii) member sequences must be aligned without long gaps. Simultaneously, the number of clusters of a given set of sequences, such as the human membrane proteome, needs to be minimized, to maximize the efficiency of the structural genomics effort.

To find the optimal clustering parameters, the domain sequences were clustered based on threshold levels of proportion of residues that can be aligned (>50%, >70%, and >90%; “coverage threshold”) and the sequence identity of the aligned segments (>20%, >25%, and >30%; “sequence identity threshold”), using program BLASTCLUST[24] (Suppl. Table 1). For quality control, the sequences in each cluster were aligned, using a multiple sequence alignment program MUSCLE[26]. We quantified the gaps in an alignment by computing its “gap ratio”, defined as the proportion of single residue gap positions in the alignment. Hence, a small gap ratio is associated with more accurate alignments that are preferred for comparative modeling. At the same coverage threshold, the gap ratio is approximately constant for the three sequence identity thresholds. As expected, the gap ratio is highest and lowest at the coverage thresholds of 50% and 90%, respectively. For further analysis, the intermediary coverage threshold of 70% was chosen.

Next, we inspected the number of members in a cluster (cluster size; Suppl. Fig. 2a). The distribution of the number of clusters as a function of cluster size (*eg*, the red bars in Suppl. Fig. 2a) is similar at the three sequence identity thresholds. Approximately 100 of the clusters contain 4 or more members. One large cluster with approximately 600 members, including the GPCR families, occurs at all three sequence identity thresholds. Finally, we assessed the current average

Nature Structural & Molecular Biology: doi:10.1038/nsmb.2508

modeling coverage of the sequences in each cluster (Suppl. Fig. 2b). Large clusters that currently have low average modeling coverage are attractive sources of targets for structural genomics, because their representative structures would result in a large increase in reliably modeled sequences.

Stage 4: Assessing target selection strategies

We assessed 'guided' and 'random' target selection strategies with respect to the number of target structures that need to be determined by experiment to achieve varying degrees of structural characterization of the human membrane proteome by comparative modeling. Guided target selection prioritizes experimental structure determination of targets by the size of the clusters that they make accessible to modeling. In contrast, random target selection picks targets randomly from a list of domain sequences without known structures.

To compare and contrast the guided and random target selection schemes, we computed the number of domain sequences that can be modeled based on the future successful determination of a 100 new target structures, with a modeling coverage threshold of 70% at different sequence identity thresholds[27]. The number of 100 was selected because it is feasible for the nine PSI centers to determine 100 membrane protein structures during two five-year cycles of PSI:Biology.

In the first scenario, we considered all unique human α-helical transmembrane domains, whether or not they can be currently modeled (Suppl. Fig. 3a; Suppl. Table 2). As expected, the guided target selection yields a significantly higher number of domain sequences covered when compared to the random target selection at the same sequence identity thresholds. For guided selection, the number of the domains that could be modeled based on 100 target structures is 1,445, 1,488, and 1,514 at 30%, 25%, and 20% sequence identity threshold, respectively (at 70% modeling coverage threshold). In contrast, for the random selection, 100 structures would allow comparative modeling of only

Nature Structural & Molecular Biology: doi:10.1038/nsmb.2508

approximately 900, 950, and 1,000 sequences at 30%, 25%, and 20% sequence identity threshold, respectively. This result highlights the superior efficiency of the guided target selection strategy over random choice[27]. Thus, the number of sequences that can be structurally characterized increases by approximately 50% when using the guided target selection strategy, as proposed here, instead of the random selection strategy.

In the second scenario, only sequences without current models were considered by first removing all clusters with sufficiently close homologs of known structure from the analysis (Suppl. Fig. 3b). As expected, guided target selection is also superior over random choice in this scenario.

The current target lists of the nine PSI membrane protein centers collectively contain 14,591 unique targets (Suppl. Table 3a). Of these, we identified 464 (3.2%) sequences that match human α-helical transmembrane domain sequences at high sequence identity (90%) and modeling coverage thresholds (90%, Suppl. Table 2). These sequences were then mapped onto the domain clusters, resulting in 57 clusters comprising a total of 1,190 sequences, which have at least one cluster member that is already included in the target lists of the nine PSI membrane protein centers. 20 of these clusters include sequences that have already been structurally characterized at the 25% sequence identity level, leaving a total of 47 clusters with 1,075 structurally uncharacterized sequences. Structure determination of at least one member from each one of these 47 clusters would thus increase the structural coverage of the human α-helical transmembrane proteome by approximately 1,000.

Stage 5: Expanding the target set by adding homologous sequences

To augment the potential target pool by the non-human homologs of the human sequences, we identified approximately 450,000 non-human sequences from approximately 50,000 different organisms homologous to the human α-helical transmembrane domains over at least 70% of the residues at the 30% sequence

Nature Structural & Molecular Biology: doi:10.1038/nsmb.2508

identity threshold. The corresponding expanded alignments can be valuable for target selection.

On average, each domain sequence has approximately 2,000 homologs from other organisms in UniProt. The number of homologs per sequence ranges from approximately 50,000 for the transmembrane domain of cytochrome b to only one for 40 transmembrane domains (Suppl. Fig. 4). For 17 human domain sequences, scanning UniProt did not identify any non-human homologs given the two thresholds used. While the non-human homologs are, as expected, predominantly from eukaryotic organisms (368,401 sequences), we were also able to identify 82,386 sequences from bacteria and 1,985 sequences from archaea (Suppl. Fig. 4).

Most of the larger clusters with a significant number of non-eukaryotic homologs have already been structurally characterized (*e.g.,* ATPases and aquaporins). Clusters with bacterial and archaeal homologs with low modeling coverage include for example SLC transporters and lipid phosphate phosphatases. A full list of the sequences with non-eukaryotic homologs, cluster sizes, and annotations can be found at salilab.org/membrane, menu item ‘Non-human Homologs’.

Nature Structural & Molecular Biology: doi:10.1038/nsmb.2508

Supplementary Note 3.

Web-based knowledgebase.

We created a dynamically updateable collaborative website (http://salilab.org/projects/membrane), using as content management system Drupal 7 (http://drupal.org). The website contains all predicted human α-helical transmembrane domain sequences, their UniProt[28] annotations, links to the associated cluster alignments, and images that indicate the modeling coverage and the location of the predicted α-helical transmembrane domains. Human α-helical transmembrane domains, the current PSI target list, and non-human homologs are linked. The results also provide clustering, comparative models when available, and a list of non-human homologous sequences. In addition, a visitor can leave comments or additions. Collaborators can edit and update the pages, as well as add pages summarizing their own data-mining efforts. These data and tools enable the PSI researchers to evaluate their current targets and potentially select additional targets, while maximizing their impact on the structural characterization of the human α-helical transmembrane proteome.

Nature Structural & Molecular Biology: doi:10.1038/nsmb.2508

Supplementary Table 1. Number of domain clusters and the corresponding gap ratio at different sequence identity and coverage thresholds.

Sequence identity threshold	20%	20%	20%	25%	25%	25%	30%	30%	30%
Coverage threshold	50%	70%	90%	50%	70%	90%	50%	70%	90%
Number of clusters	1,059	1,186	1,394	1,074	1,201	1,406	1,106	1,236	1,435
Gap ratio	11.3%	8.2%	5.5%	10.9%	8.0%	5.4%	10.4%	7.7%	5.4%

A small gap ratio is associated with more accurate alignments that are preferred for comparative modeling. For the current analysis, the gap ratio was computed for the alignments obtained at different sequence identity thresholds and coverage thresholds.

Nature Structural & Molecular Biology: doi:10.1038/nsmb.2508

Supplementary Table 3: Target lists for 9 PSI membrane protein centers.

a) Number of PSI targets that match human α-helical transmembrane domains.

Center	Total number of targets	Number of targets that match transmembrane domains (% sequence identity / % residues aligned)					
		25/50	25/90	30/50	30/90	90/50	90/90
CSMP[a]	2,454	835	362	509	258	106	96
GPCR[b]	127	127	126	127	126	127	125
MPID[c]	85	8	1	2	0	0	0
MPSBC[d]	40	31	8	18	2	0	0
MPSbyNMR[e]	12	6	6	5	4	4	4
NYCOMPS[f]	12,881	4,840	1,599	2,859	1,041	232	218
TEMIMPS[g]	89	30	18	26	17	9	8
TMPC[h]	112	43	28	36	24	8	8
TransportPDB[i]	586	563	549	561	535	510	499
Total (unique sequences)	14,591	6,264	2,600	4,003	1,942	965	928

b) Number of human α-helical transmembrane domains that match PSI targets.

Center	Number of human α-helical transmembrane domains that match PSI targets (% sequence identity / % residues aligned)					
	25/50	25/90	30/50	30/90	90/50	90/90
CSMP[a]	1,581	892	1,016	613	104	96
GPCR[b]	873	796	813	749	180	176
MPID[c]	12	1	2	0	0	0
MPSBC[d]	96	11	28	4	0	0
MPSbyNMR[e]	146	111	54	42	4	4
NYCOMPS[f]	2,210	1,146	1,354	761	96	87
TEMIMPS[g]	221	141	112	82	10	9
TMPC[h]	374	281	201	142	10	10
TransportPDB[i]	950	770	856	757	748	719
Total (unique sequences)	3,017	2,216	2,563	1,991	1,086	1,039

[a]Center for Structure of Membrane Proteins, [b]GPCR Network, [c]Center for Membrane Proteins in Infectious Diseases, [d]Membrane Protein Structural Biology Consortium, [e]Membrane Protein Structures by Solution NMR, [f]New York Consortium on Membrane Protein Structure, [g]Transcontinental EM Initiative for Membrane Protein Structure, [h]Transmembrane Protein Center, [i]Center for the X-ray Structure Determination of Human Transporters. Suppl. Table 2a shows the actual PSI targets and their similarity to the closest human α-helical transmembrane domain(s). Suppl. Table 2b show how many human α-helical transmembrane domains are related to these targets. The PSI target sequences were retrieved from TargetTrack (Oct 2012), the structural biology target registration database (http://sbkb.org/tt/).

Nature Structural & Molecular Biology: doi:10.1038/nsmb.2508

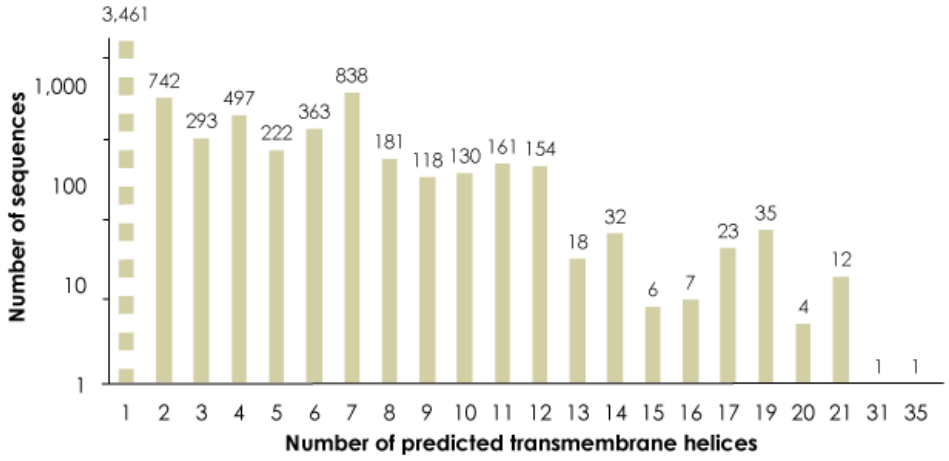
Supplementary Table 2. Guided *versus* random target selection*.

Trans membrane domains	Sequence identity threshold (%)	20%	25%	30%
	Number of domains	2,925	2,925	2,925
	Number of domains without models	1,932	2,642	2,674
Guided Target Selection	Number of domains (100)	1,514	1,488	1,445
	Number of domains without models (100)	793	1,407	1,348
	Number of modeled domains (100)	**1,786**	**1,690**	**1,599**
Random Target Selection	Number of domains (100)	1,000	950	900
	Number of domains without models (100)	400	930	920
	Number of modeled domains (100)	**1,037**	**1,213**	**1,171**

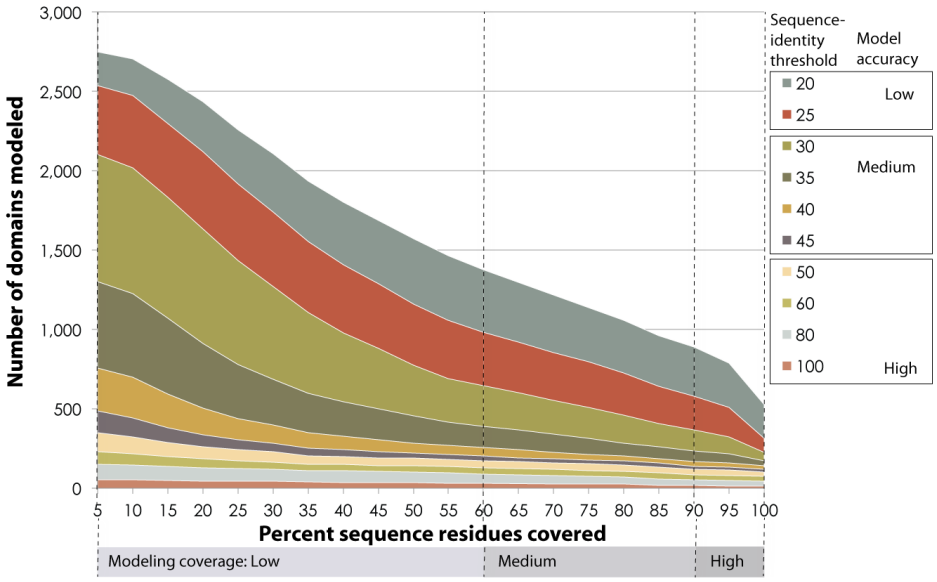
* The guided and random target selection strategies are assessed with respect to structural coverage. For both strategies, 100 target structures have been assumed (Stage 4) and the 70% coverage threshold was imposed. For guided (random) target selection: 'Number of domains (100)', the number of human α-helical transmembrane domains in the 100 largest (random) domain clusters. 'Number of domains without models (100)', the number of transmembrane domains in the 100 largest (random) domain clusters with low or medium model coverage. 'Number of modeled domains (100)', the number of modeled domains after determining the 100 target structures. The number of domains in the first 100 clusters increases when the clustering sequence identity threshold decreases from 30% to 25%, because the number of sequences per cluster increases. The number of domains without models decreases when the sequence identity threshold drops to 20%, because the sequences in the largest cluster (the GPCR sequences) become modeled.

Nature Structural & Molecular Biology: doi:10.1038/nsmb.2508

a)

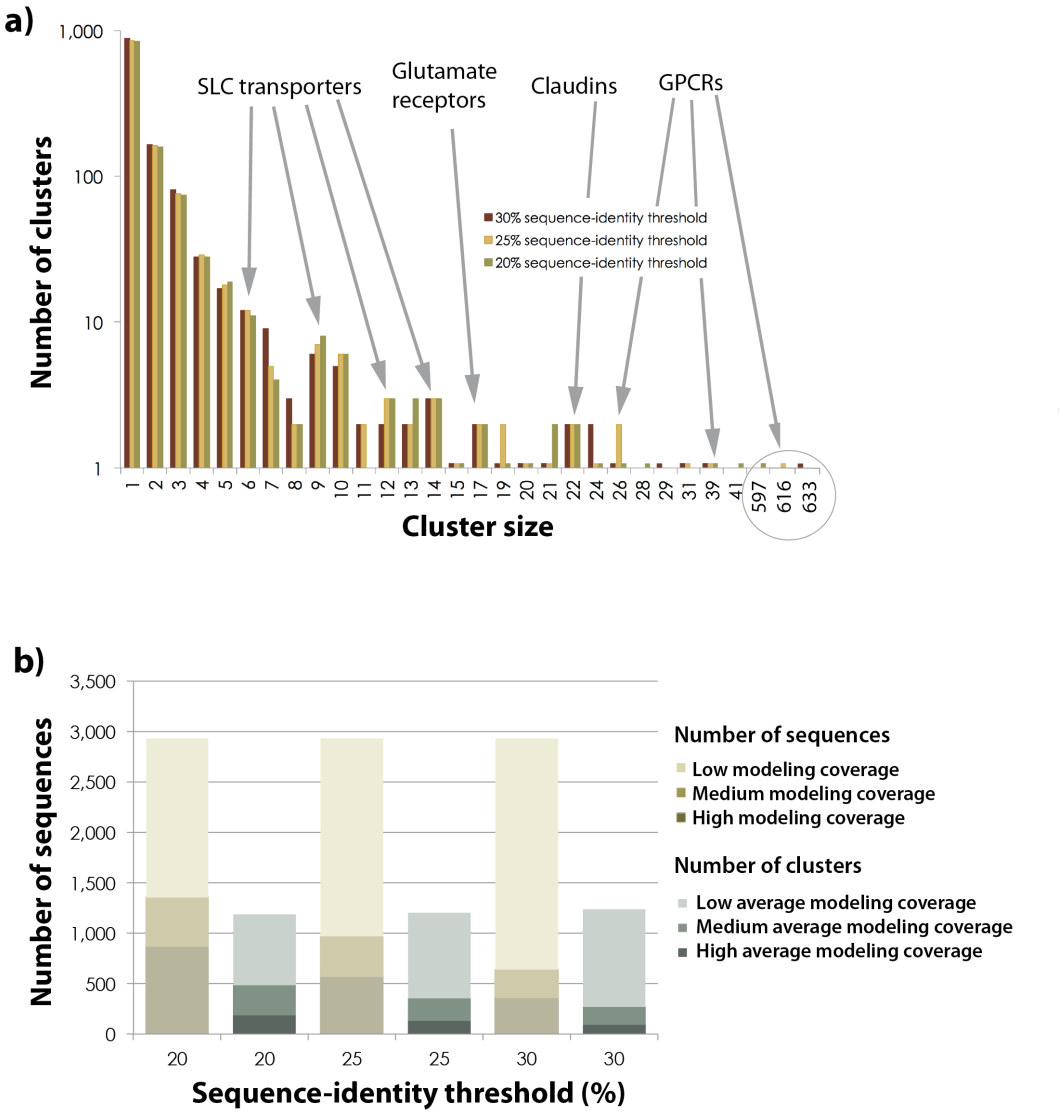


b)



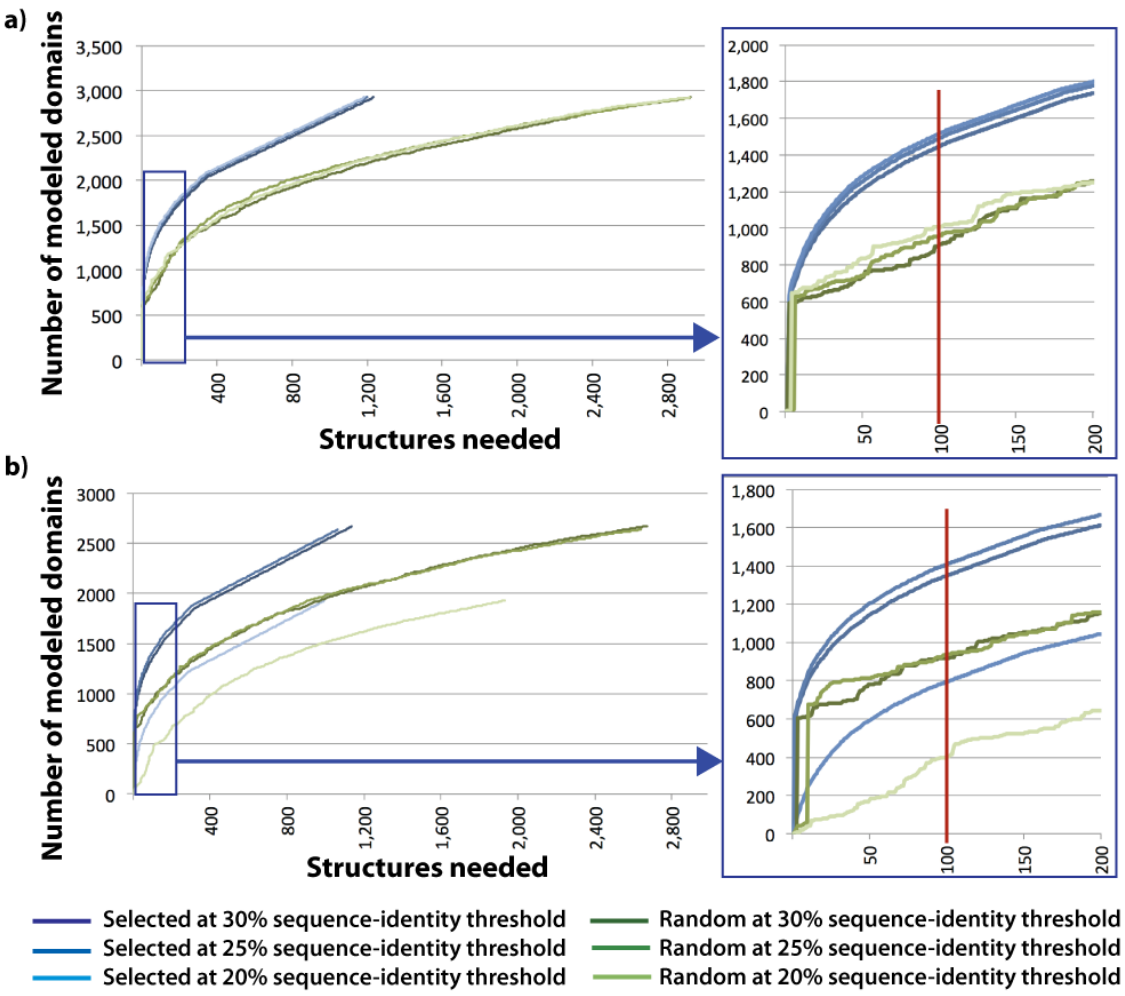
Supplementary Figure 1. **a)** The number of sequences with a given number of predicted transmembrane α-helices in human transmembrane domains. The number of sequences with one predicted transmembrane α–helix is indicated by a hashed bar; these sequences have not been included in the subsequent analysis. **b)** Modeling the human α-helical transmembrane domain proteome based on currently known structures. The number of modeled domains is shown as a function of the modeling coverage threshold (percent sequence residues covered), for the sequence identity thresholds ranging from 20% to 100%. The number of domains that can be modeled increases with the lower sequence identity and lower model coverage thresholds.

Nature Structural & Molecular Biology: doi:10.1038/nsmb.2508



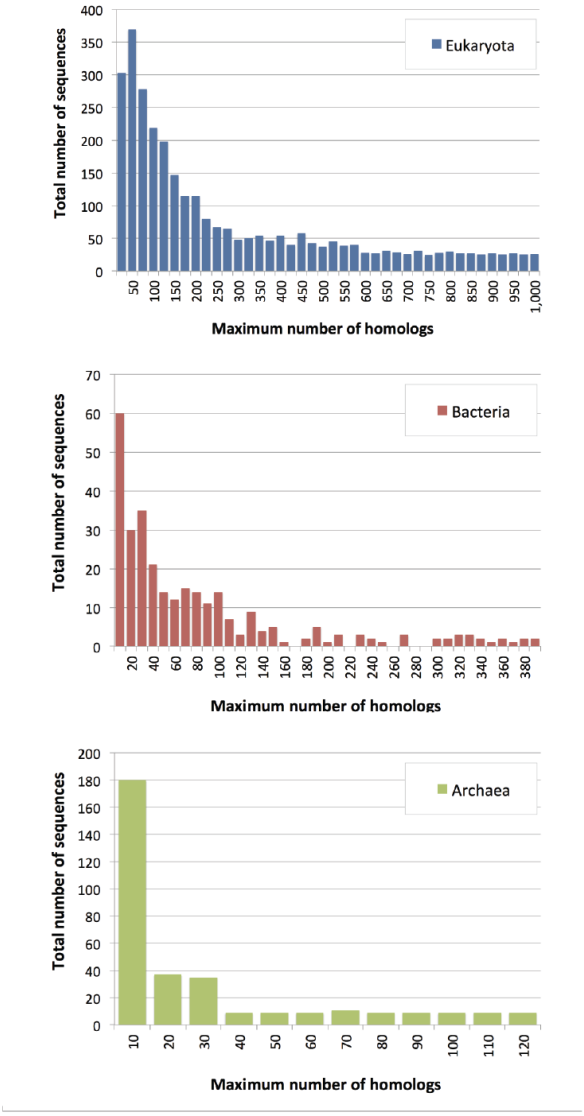
Supplementary Figure 2. Clustering. a) Distributions of cluster sizes for 30%, 25%, and 20% sequence identity thresholds. The clusters of several superfamilies are indicated. b) The number of sequences and clusters for varied modeling coverage and sequence identity thresholds, used both for clustering and calculating the modeling coverage) at 70% coverage threshold for the clustering.

Nature Structural & Molecular Biology: doi:10.1038/nsmb.2508



Supplementary Figure 3. Target selection schemes. The number of sequences modeled is shown for six scenaria, corresponding to the random and guided target selections at three sequence identity thresholds. The plots on the right zoom into the plots on the left, for more detail. **a)** All clusters are considered, including those with homologs of already determined membrane protein structures. **b)** Already known membrane protein structures are taken into account by removing clusters with at least one sequence for which at least 90% of residues are modelable at the given sequence identity threshold.

Nature Structural & Molecular Biology: doi:10.1038/nsmb.2508



Supplementary Figure 4. Expansion of human α-helical transmembrane domain clusters by their homologs from other species. The total number of human domain sequences (y-axis) with a given maximum number of non-human homologs (in bins) in UniProt is shown for homologous sequences from eukaryotes, bacteria, and archaea. For clarity, the x-axis is truncated (there are 745 sequences with up to 50,000 homologs from eukaryotes, 128 sequences with up to 7,000 homologs for bacteria, and 6 sequences with up to 126 homologs for archaea).

Nature Structural & Molecular Biology: doi:10.1038/nsmb.2508

Supplementary Note 4: References for Supplementary Materials

1. Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E.L. *Journal of molecular biology* 305, 567-80 (2001).
2. Fagerberg, L., Jonasson, K., von Heijne, G., Uhlen, M. & Berglund, L. *Proteomics* 10, 1141-9 (2010).
3. Nugent, T. & Jones, D.T. *PLoS computational biology* 6, e1000714 (2010).
4. Kall, L., Krogh, A. & Sonnhammer, E.L. *Journal of molecular biology* 338, 1027-36 (2004).
5. Kall, L., Krogh, A. & Sonnhammer, E.L. *Bioinformatics* 21 Suppl 1, i251-7 (2005).
6. Bernsel, A. et al. *Proceedings of the National Academy of Sciences of the United States of America* 105, 7177-81 (2008).
7. Viklund, H., Bernsel, A., Skwark, M. & Elofsson, A. *Bioinformatics* 24, 2928-9 (2008).
8. Zhou, H. & Zhou, Y. *Protein science : a publication of the Protein Society* 12, 1547-55 (2003).
9. Larsson, T.P., Murray, C.G., Hill, T., Fredriksson, R. & Schioth, H.B. *FEBS Lett* 579, 690-8 (2005).
10. Edgar, R.C. *Bioinformatics* 26, 2460-1 (2010).
11. Eswar, N. & Sali, A. in *From Molecules to Medicine, Structure of Biological Macromolecules and Its Relevance in Combating New Diseases and Bioterrorism* (eds. Sussman, J.L. & Spadon, P.) 139-151 (Springer-Verlag, Dordrecht, The Netherlands, 2009).
12. Pieper, U. et al. *Nucleic Acids Research* 39, D465-74 (2011).
13. Sanchez, R. & Sali, A. *Proc Natl Acad Sci U S A* 95, 13597-13602 (1998).
14. Altschul, S.F. et al. *Nucleic acids research* 25, 3389-402 (1997).
15. Sali, A. & Blundell, T.L. *J Mol Biol* 234, 779-815 (1993).
16. Smith, T.F. & Waterman, M.S. *Journal of molecular biology* 147, 195-7 (1981).
17. Eswar, N. et al. *Current Protocols in Bioinformatics* Chapter 5, Unit 5.6 (2006).
18. Marti-Renom, M.A., Madhusudhan, M.S. & Sali, A. *Protein Sci* 13, 1071-1087 (2004).
19. Dutta, S., Zardecki, C., Goodsell, D.S. & Berman, H.M. *Journal of applied crystallography* 43, 1224-1229 (2010).
20. Shen, M.Y. & Sali, A. *Protein Sci* 15, 2507-2524 (2006).
21. Melo, F., Sanchez, R. & Sali, A. *Protein Sci* 11, 430-448 (2002).
22. Eramian, D. et al. *Protein Sci* 15, 1653-1666 (2006).
23. Sanchez, R. & Sali, A. *Methods Mol Biol* 143, 97-129 (2000).
24. Johnson, M. et al. *Nucleic acids research* 36, W5-9 (2008).
25. Chen, L., Oughtred, R., Berman, H.M. & Westbrook, J. *Bioinformatics* 20, 2860-2 (2004).
26. Edgar, R.C. *BMC bioinformatics* 5, 113 (2004).
27. Vitkup, D., Melamud, E., Moult, J. & Sander, C. *Nat Struct Biol* 8, 559-66 (2001).
28. *Nucleic acids research* 40, D71-5 (2012).

Nature Structural & Molecular Biology: doi:10.1038/nsmb.2508